Practical Assessment, Research, and Evaluation

Volume 13 *Volume 13, 2008* Article 9

2008

A comparison of two different methods for setting performance standards for a test with constructed-response items

Gunilla Näsström

Peter Nyström

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 13, Number 9, September 2008 ISSN 1531-7714

A comparison of two different methods for setting performance standards for a test with constructed-response items

Gunilla Näsström and Peter Nyström, *Umeå University, Sweden*

The trustworthiness of performance standards influences the credibility of criterion-referenced large-scale testing. In this paper, two standard-setting methods are evaluated and compared, when applied to a test with polytomously scored constructed-response items. A version of the Angoff method is chosen as representative of the class of test-centred standard-setting procedures and the borderline-group method represents the class of examinee-centred procedures. The evaluation is based on procedural, internal and external evidence. The results indicate that both methods provide reasonable and trustworthy approaches to standard setting, but also confirm some of the potential problems with these methods.

Inferences from criterion-referenced large-scale testing rely heavily on the credibility of the thresholds used to indicate whether a student performance meets a certain standard or not. These thresholds, or performance standards, are estimated in a process called standard-setting and often defined as positions on the score scale (cut-scores). There are no true, objective or "golden" performance standards for any assessment (Kane, 1998a), and the performance standards can only be set in a more or less trustworthy way. To achieve credible performance standards, a large number of methods have been proposed. The different standard setting methods are well researched for tests entirely made up of selected-response items, but for tests with constructed-response items the research is much more sparse (Hambleton & Pitoniak, 2006).

This study concerns performance standards, which can be viewed as operationalizations of learning objectives on an assessment indicating if the examinees have achieved a sufficient level of knowledge and/or skills (Hambleton & Pitoniak, 2006). These performance standards are composed of performance levels, performance descriptions and cut-scores (Hansche, 1998). Performance levels are labels for specific levels of performance, for example below basic, basic, proficient and advanced used in National Assessment of Educational Progress (Kane, 1998b) and fail, pass, pass with distinction and pass with special distinction used in national tests in Sweden (Skolverket, 2005). Performance descriptions are narrative descriptions of how well examinees should perform at each performance level (Hansche, 1998). A cut-score is a point on the score scale for a particular test associated to a performance level (Kane, 2001) and divides the examinees into two performance categories based on their performance on the particular assessment (Cizek & Bunch, 2007).

Standard-setting methods

The large number of methods for setting performance standards described in the literature (see e.g. Cizek & Bunch, 2007) can generally be characterized as examinee-centred, test-centred or a combination of these two approaches (Jaeger, 1989). Which method to choose depends on the advantages and disadvantages of different methods in different contexts. Kane (1994) proposed three types of evidence that should be supplied in order to defend the performance standards set using a chosen method. These are procedural, internal, and external evidence. The three types of evidence will be further elaborated later in this paper, and used in the evaluation of methods.

Examinee-centred methods

Examinee-centred methods are based on judgments about examinees. In examinee-centred methods judges categorize examinees according to performance level

(e.g. non-qualified, qualified and borderline) based on some external criterion other than the test score (Giraud, Impara & Buckendahl, 1999/2000). Typically, the test is then administrated to the categorized examinees and the cut-score is set based on their results on the test (Cizek, 2006). The two most common examinee-centred methods are the borderline-group method and the contrasting group method (see e.g. Hambleton & Pitoniak, 2006). The borderline-group method is chosen as an example of the examinee-centred methods in this study, primarily because it is regarded as conceptually simple (Jaeger, 1989; Hambleton & Pitoniak, 2006) and recommended for holistic and constructed-response tests (Kane,1998a).

In the borderline-group method, judges are asked to conceptualize the characteristics of border-line examinees and identify specific examinees that fit these characteristics (Livingstone & Zieky, 1982). Then the assessment is administrated, scored and analysed and the median score of those defined as borderline examinees is typically used as the cut-score (Cizek, 2006). If there are more than one cut-score to be set, a borderline group for each cut-score has to be identified (Cohen, Kane & Crooks, 1999). According to Hambleton, Jaeger, Plake & Mills (2000) the borderline-group method is group-dependent, which means that if the sample of examinees and judges are different from the distribution of the whole population, then the credibility of the cut-scores can be questioned. However, identifying "truly" borderline examinees is more important than having representative samples (Livingstone and Zieky, 1982).

Advantages of the borderline-group method are the conceptual simplicity of the method (Hambleton & Pitoniak, 2006), and the fact that the judges deal with familiar individual examinees (Livingstone & Zieky, 1982). Disadvantages of the borderline-group method are that the method is time-consuming (Kane, 1998a), and requires a large panel of judges (Hambleton & Pitoniak, 2006) and a large sample of examinees (Cizek, 2006). There is also a tendency for judges to include factors and performances not covered by the assessment in the categorization of examinees, (Hambleton et al., 2000) and to identify examinees as borderlines when there is uncertainty about their performance (Jaeger, 1989; Hambleton & Pitoniak, 2006). A potential problem with the borderline-group method is that the cut-score arrived at by teachers with high-performing examinees tends to be higher than the cut-scores from teachers with lower-performing classes (Livingstone & Zieky, 1989).

Test-centred methods

Test-centred methods are based on judgments about the items in a particular assessment. During the review of the assessment items, the judges decide on the level of performance required to meet each performance standard (Kane, 1998a). This is done by judgments about expected performance on each item for hypothetical examinees just barely fulfilling the requirements for a certain performance standard (Hambleton & Pitoniak, 2006). The Angoff method, Ebel's procedure, Jaeger's method, the Nedelsky procedure and the Bookmark method are well-known examples of test-centred methods, which have been modified and extended in many ways (Kane, 1998a; Hambleton & Pitoniak, 2006). The Angoff method is chosen to represent the test-centred methods in this study because in its original version, or in a modified and extended version, it is the most widely used procedure for standard-setting (Hurtz & Auerbach, 2003). Furthermore, a modified version of the Angoff method is used regularly as the standard setting procedure for the national tests in mathematics in Sweden.

When the Angoff method is applied to tests with items scored as right or wrong, the judges are asked to conceptualize a group of just barely qualified examinees and to estimate the proportion of this group which would answer each item in the test correctly (Cizek, 2006). For each judge the estimated probabilities are summed and these sums are averaged across judges to arrive at a recommended cut-score (Ferdous & Plake, 2007). For tests with polytomous scored items, the average proportion of full credit is estimated for the barely qualified examinees for each item. A recommended cut-score is calculated by multiplying these estimates by the maximum score of each item, summarising the products, and averaging across judges.

The advantages of the Angoff method are that it is easy to administrate, that it gives compensatory cut-score (i.e. a high score on one item can balance a low score on another item (Hambleton & Pitoniak, 2006)), and that the method can be implemented before the administration of the test (Kane, 1998a). Disadvantages are the atomistic nature of the method (Hambleton et al., 2000), the difficulty for the judges to estimate the performance on individual items for a group of just barely qualified examinees, and the tendency to

overestimate performance on easy items and underestimate difficult items (Hambleton & Pitoniak, 2006).

Aim

The aim of this study is to compare the validity of two different methods for determining cut-scores on a Swedish national test in mathematics.

More specifically we want to

- evaluate the trustworthiness of cut-scores resulting from a test-centred and an examinee-centred approach to standard-setting, and
- compare the inferences of the different cut-scores with respect to the distribution of examinees over performance levels.

METHOD

The features and consequences of test-centred versus examinee-centred procedures for standard-setting are studied in the context of a national test in mathematics. Performance standards for the test were set with two different methods, an Angoff procedure and a borderline-group procedure. The inferences of the cut-scores resulting from the two different procedures are evaluated in different ways, partly based on the application of these cut-scores on a large, nation-wide, sample of student results.

The study is based on a Swedish national test in mathematics given in spring 2004. The test consists of 22 items, all constructed-response. Student responses were scored dichotomously for 9 of the items, and polytomously for 13 of the items (from 2 to 6 points each). The maximum score on the test is 40. Two potential cut-scores are evaluated here, one for the performance level Pass (P) and one for the performance level Pass with distinction (PD)

Standard setting procedures

For the borderline-group procedure, the judges were initially sampled from a pool of teachers engaged in the development of national tests in mathematics. The process of developing national tests in Sweden involves many teachers who through this work acquire familiarity with the national objectives and performance descriptions. In addition they are well acquainted with the general structure of the national tests. In this way teachers representing 20 schools were asked to participate in the study. In order to reach the goal of at least 100 examinees in each borderline group, which was considered a minimum for arriving at reliable cut-scores, eight other schools were selected.

All of the 28 schools were invited to participate with up to six teachers, where large schools were encouraged to participate with more teachers than small schools, and 24 schools participated in the study. Complete and useful records were reported from 44 teachers predicting the performance of 46 groups of examinees, 948 examinees in all. The participating teachers had at least one group of examinees who were going to take the particular test. In the sample of participating teachers there were as many women as men. Most of the teachers were very experienced. As many as 36 of the teachers had at least 6 years of experience as teachers in upper secondary schools, 2 teachers had between 3 and 5 years of experience and 6 teachers had up to 2 years of experience.

Approximately one month before the national test the teachers predicted the performance of their examinees on the coming national test, without seeing the actual test. For the examinee-centred method it is important that the judges categorize their examinees based on skills defined by the test specifications, instead of their expected performance on the test items (Giraud et al. 1999/2000).

The scale for prediction was based on three of the grades used in upper secondary schools in Sweden: Fail (F), Pass (P), and Pass with distinction (PD). To nuance the scale, teachers often use + and – together with the grades when they discuss grades during the course. Teachers are used to this way of constructing a more fine-grained scale, indicating relative performance within a basically (theoretically) criterion-referenced grade system. The borderline group for P was constructed by combining the groups of examinees on the scale steps F+ and P-, and for PD the scale steps P+ and PD- were combined. [1] The median value of the borderline examinees' test results was calculated and used as the resulting cut-score.

The Angoff-procedure is described in Lindström (2003). In our study, a panel of 11 mathematics teachers (4 female and 7 male) was appointed. The panelists had

[1] Actually a forth grade was part of the teachers' prediction, Pass with special distinction. However, for the sake of this study, this category was included in the highest nuance of Pass with distinction (PD+).

more than five years of teaching experience and had taught the specific mathematics course that the national test was assessing. Due to their experience, the teachers were expected to be well acquainted with the national objectives and performance descriptions. Each of the panel members had participated in similar panels performing the Angoff procedure at least three times prior to this occasion. For the performance level Pass the panelists discussed their Angoff estimates and then made a new estimation. In this iterative version of the Angoff method, the resulting cut-score is based on the mean-values from the second step.

The procedure is a two-step, iterative extended Angoff-method. In the first step, the judges individually estimate the performance on each item for a group of just barely qualified examinees for each performance level. This first step is followed by a discussion between the judges about differences in their estimations. A second, similar step of estimations takes place after the discussion. The resulting cut-scores are based on the mean-values from the second step.

Validity evidence

The analysis of the results is based on the three kinds of validity evidence proposed by Kane (1994): procedural, internal and external. *Procedural evidence* deals with how reasonably, systematically and defensibly the standard setting procedure has been carried out. *Internal evidence* deals with data generated within the standard-setting procedure and with a special focus on consistency of the results. A common rule-of-thumb applicable to the Angoff method is that low standard deviations between judges indicate high inter-judge consistency and high confidence of the resulting cut-scores (Hambleton & Pitoniak, 2006). For the borderline-group procedure Livingstone and Zieky (1989) argued that the derived cut-scores are trustworthy, if the scores of each borderline group show small standard deviations and if their mean scores are ordered. *External evidence* is based on comparisons with external sources, e.g. other measurements of the same knowledge and/or skills, results from other standard-setting procedures, and group distribution when the test is given. A cut-score is viewed as more trustworthy if different standard-setting procedures result in similar performance standards (Hambleton & Pitoniak, 2006). In this study, the evaluation of external evidence is based on teachers' reports of results from the national tests and students final course grades.

Cut-scores resulting from the standard-setting procedures were applied to the reported results of a national sample of examinees (n=6561), and the resulting grade-distributions were analysed and compared to the distribution of final course-grades.

In addition, since the item difficulties for examinees performing at the cut-scores set by the Angoff procedure can be seen as the test-use equivalence of Angoff estimates, these values can be compared as an evaluation of the Angoff method. For this purpose, average proportions of full credit were calculated for each item for examinees performing at different cut-scores. The distribution of these p-values over the total score of the test was modeled using a two-parameter logistic model $P_i(S_j)=\frac{e^{a_i(S_j-b_i)}}{1+e^{a_i(S_j-b_i)}}$, where P is the probability that a student j with the total score Sj will answer the item i in a way that gives full credit (Lindström, 2003)[2]. When the parameters ai (a measure of the discriminating power of the item) and bi (a measure of the item difficulty) have been estimated, P can be calculated for values of S equal to the cut-scores set through the Angoff procedure. These values were compared to the Angoff estimates.

RESULTS

Results from the standard setting procedures

In the borderline-group procedure the judges were initially asked to predict their examinees' performances on the national test. The result of this categorisation is shown in Table 1. The group PD+ is large because it contains examinees performing at a higher level (Pass with special distinction) and this study only focuses on three of the four grades on the prediction scale.

Based on these categorizations, five groups were formed, including two borderline groups. Borderline group 1 consists of those examinees who were predicted to perform in the F+ to P- interval (n = 123). Similarly, borderline group 2 was constructed as the examinees who were expected to perform in the P+ to PD- interval (n = 213). Test-results were analysed with regard to these groups (see Table 2).

[2] This model has been used for item analysis in the development of Swedish national tests at the Department of Educational Measurement and has been proven useful and valid, e.g. for analysis of differential item functioning.

Following the recommendation to use the median score from the borderline procedure as the final performance standard, the cut-score for P will be 9 and the cut-score for PD will be 19.

In this study, the distributions of score points for borderline-group 1 and borderline-group 2 have standard deviations of 5.88 and 6.53 respectively (see Table 2). The score distribution of the total-group of examinees has a standard deviation of 9.65. This means that the standard deviation of borderline-group 1 was 61% of the total-group standard deviation and 68 % for borderline-group 2.

Table 1 Teachers' predictions of their examinees' performances on the test. The predictions are based on the broader categories of Fail (F), Pass (P), Pass with distinction (PD) and Pass with special distinction (PSD), with – and + indicating low and high performances within each category. All examinees predicted as PSD are added to group of PD+.

Grade	Subcategory	N	
Fail	F-	12	F-group
	F	74	
	F+	39	Borderline group 1
	P-	84	
Pass	P	165	P-group
	P+	133	Borderline group 2
	PD-	80	
Pass with distinction	PD	161	PD-group
	PD+	200	
	Total	948	

Table 2. The examinees' test results in each prediction group

	Total score					
Category of examinees	N	Mean	Standard deviation	1st quartile	Median	3rd quartile
F-group	86	6.33	4.76	3	6	9
Borderline-group 1	123	9.72	5.88	5	9	14
P group	165	15.09	6.09	10	15	19
Borderline-group 2	213	18.62	6.53	14	19	24
PD-group	361	27.45	6.30	24	28	28
Total	948	19.10	9.65	12	19	27

A potential problem with the borderline-group method is that the cut-score arrived at by teachers with high-performing examinees tend to be higher than the cut-scores from teachers working with lower-performing classes. In Figure 1 median performances of the examinees in borderline-group 1 are plotted against the median performances of the whole class, for classes with at least four examinees in borderline-group 1. The relationship is clearly positive, indicating that teachers of high-performing classes tend to define high-performing borderline-groups, which results in higher cut-scores.

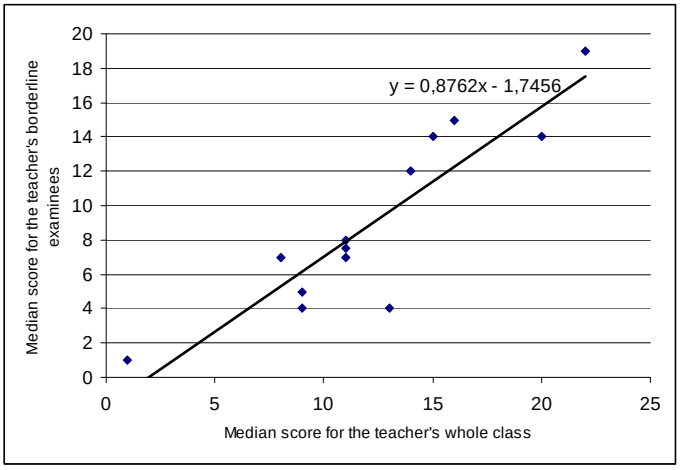

Figure 1: The relation between median scores for examinees in borderline-group 1 and the median score for all of the examinees taught by different teachers.

A similar relationship is found for borderline-group 2 (see Figure 2).

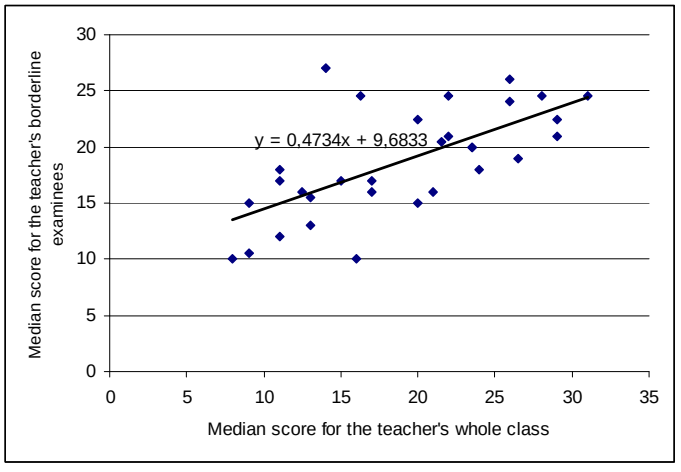

Figure 2: The relation between median scores for examinees in borderline-group 2 and the median score for all of the examinees taught by different teachers.

The results from the standard setting procedure using the Angoff method are presented in Table 3. The resulting cut-scores from the Angoff procedure were 10 for P and 22 for PD (see Table 3). The median values are about the same as the mean values.

Application of the cut-scores on a national sample of examinees

In comparison, the Angoff procedure gave more demanding cut-scores than the borderline procedure. The difference was more pronounced for the performance level PD than for P. The results of applying the different cut-scores to a national sample of student performances are presented in Table 4.

Self-evidently, using the higher cut-scores from the Angoff procedure results in fewer examinees achieving higher performance levels. For 5671 examinees (87 %) the inferences from the test-results are the same for the two different standard setting procedures, i.e. 13 % of the categorisations differ between the standard-setting procedures. Out of the 1578 test-results that were categorised as F based on the Angoff cut-score, 228 (14 %) received a higher test-grade using the cut-score from the borderline procedure. Similarly, 24 % of the results categorised as P using the Angoff cut-score attained a higher test-grade using the borderline-group cut-score.

Table 5 presents a comparison between course-grades set by teachers and test-grades resulting from applying the cut-scores from the borderline-group procedure to results from the national sample. For the borderline-group procedure, course-grades and performance levels on the test have a 79 % agreement. The correlation between these two measures is 0.81.

In Table 6, the examinees' performance levels on the test, with cut-scores set by the Angoff procedure, are compared to the examinees' course grades set by their teachers. For 80 % of the examinees, the course-grade corresponded to the performance level indicated by the test using the cut-score from the Angoff procedure. The correlation between these two measures is 0.83.

A comparison between Tables 5 and 6 indicates that the teachers' course grades correspond more with the cut-scores from the borderline-group procedure for F and P, and more with the cut-scores from the Angoff procedure for PD.

Table 3. The judges' finial estimation in the extended Angoff procedure for the cut-scores at the two performance levels (N = 11)

Performance standard	Total score				
	Mean	Standard deviation	1st quartile	Median	3rd quartile
Passed (P)	10.22	1.13	9.65	10.00	10.75
Passed with distinction (PD)	21.80	2.35	20.30	22.10	23.15

Table 4. The number of examinees in the national sample who will attain the three performance levels on the test based on the cut-scores from each of the standard setting procedures.

Angoff procedure	Borderline-group procedure			
	F	P	PD	Total
F	1350	228	0	1578
P	0	2101	662	2763
PD	0	0	2220	2220
Total	1350	2329	2882	6561

Table 5. Comparison between course grades set by teachers and the performance levels on the national test based on cut-scores from the borderline-group procedure. Number of examinees in parenthesis.

		Course grade			No. of examinees
		F	P	PD	
Performance level on the test	F	73% (990)	26% (354)	0% (6)	1350
	P	10% (239)	81% (1876)	9% (214)	2329
	PD	0% (6)	20% (574)	80% (2302)	2882
No. of examinees		1235	2804	2522	6561

Table 6. Comparison between course grade set by teachers and the performance levels on the national test with cut-scores set by the Angoff procedure. Number of examinees in parenthesis.

		Course grade			No. of examinees
		F	P	PD	
Performance level on the test	F	69% (1088)	31% (483)	0% (7)	1578
	P	5% (144)	77 % (2137)	17% (482)	2763
	PD	0% (2)	8 % (184)	92% (2033)	2220
No. of examinees		1235	2804	2522	6561

The basis of the Angoff procedure is that for each item in a test, the judges estimate the achievement of the student barely passing a certain performance standard. These estimates can be compared to the actual performance of examinees at the particular cut-score set by the Angoff procedure.

In Figure 3, p-values for students performing at the cut-score suggested by the Angoff procedure are plotted against Angoff estimates, for each item.

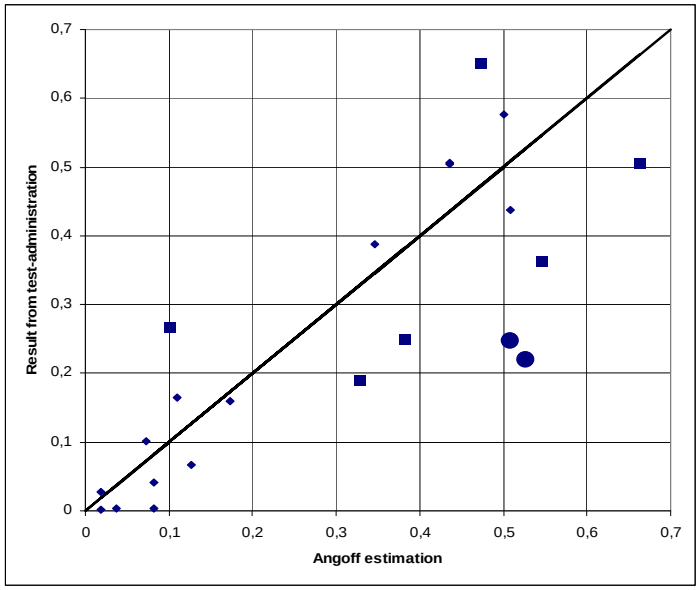

Figure 3. Item correlation between estimated p-values in the Angoff procedure and actually p-values when using the cut-score derived by the Angoff procedure, for the performance level P.

Small squares in Figure 3 indicate items where the empirically found p-values at the cut-score were within one standard deviation from the Angoff estimates. The larger squares indicate items where the deviation was more than one but less than two standard deviations away from the Angoff mean estimates and the filled circles indicate items where the test results were more than two standard deviations. The standard deviations refer to the variance of the judges' Angoff estimates around the mean.

For the cut-score for P, only two Angoff estimates deviated more than two standard deviations from the empirically found p-values. If the significance demand is lowered to one standard deviation, another six deviations are identified. Out of the total of eight significant deviations, six were over-estimations and two were under-estimations.

For the cut-score for PD (see Figure 4), 12 Angoff estimates deviated more than one standard deviation from empirically found p-values and five of these deviated more than two standard deviations. Out of the 12 significant deviations, eight were over-estimations and four were un-der-estimations.

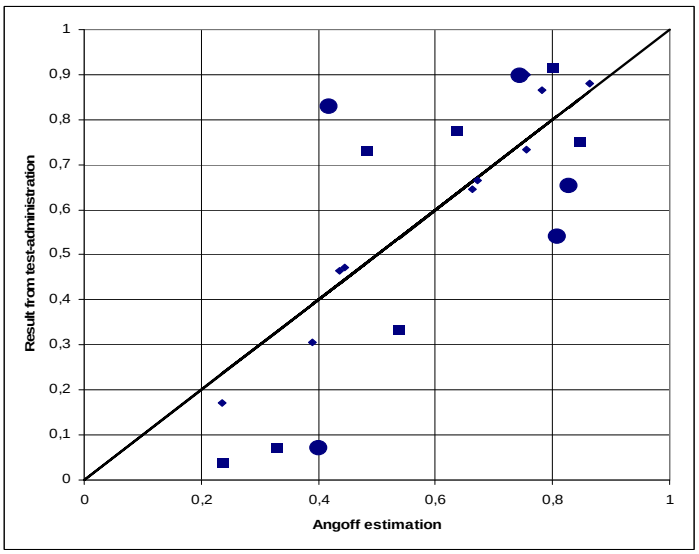

Figure 4. Item correlation between estimated p-values in the Angoff procedure and actually p-values when using the cut-score derived by the Angoff procedure, for the performance level PD.

ANALYSIS AND DISCUSSION

The purpose of this study is to compare the validity of two different methods for determining cut-scores. Specifically, the trustworthiness of cut-scores derived from a test-centred (the Angoff procedure) and an examinee-centred (borderline-group procedure) approach to standard setting is evaluated and the inferences of the different cut-scores with respect to the distributions of examinees are explored. The results are analysed and discussed with respect to the three kinds of validity evidence suggested by Kane (1994), i.e. procedural, internal and external evidence.

Procedural evidence

Procedural evidence concerns how the standard setting procedures were carried out, and in our case both procedures followed most of the important steps recommended in the literature. One exception was that none of the procedures included any training of the judges or evaluation by the judges. Training of the judges

is regarded as an important part of any standard setting procedure (Hambleton & Pitoniak, 2006), and the purpose is to give the judges, for the specific method, the necessary skills, which is not feasible in the selection of the judges (Raymond & Reid, 2001). However, the background of the judges in this study, their prior experience of standard-setting and/or work with national tests makes it unlikely that further training would have made any major difference for the outcome of the standard-setting procedures. The judges involved in the Angoff method all had prior experience of the procedure. In addition, an iterative step was included, when judges were given the opportunity to evaluate their initial judgments based on the judgments made by the other judges. However, the changes made by judges are rare and generally small (results not shown). The panelists were not given impact data as part of the Angoff procedure because the test was not yet administered to students and the pretesting data was not representative enough to be possible to use for this purpose. The use of impact data could improve the results from the Angoff procedure, but requires substantial changes of the pretest procedure which are not easily accomplished.

In the borderline-group procedure a necessary skill is to be able to categorize examinees according to expected performance on a test. Teachers in Sweden assess regularly and have full responsibility for grading their students, which gives them experience of categorizing students within a grade scale. The same grade scale is used for the national tests, the purpose of which is to support the teachers in their grading of students. Therefore, teachers working in schools are experienced in categorizing students on the grade scale and acquire the skill necessary for judges in the borderline-group procedure. In this study, most of the teachers had at least two years of teaching experience and are therefore assumed to have the necessary skill to be judges in the borderline-group procedure.

Another exception was that only one panel of judges was used to the Angoff procedure. The trustworthiness of the Angoff procedure is enhanced by using more than one panel.

Internal evidence

Internal evidence deals with data generated within each standard-setting procedure, with a special focus on consistency of the results. Smaller standard deviations indicate higher inter-judge consistency and therefore higher trustworthiness in the derived cut-scores (Hambleton & Pitoniak, 2006, Livingstone and Zieky, 1989). The standard deviations for the borderline-groups were 61 % and 68 % of the total-group standard deviation, which are low compared to the 86 % found in the study by Livingstone and Zieky (1989). With the smaller proportions of the total-group standard deviations for the borderline-groups in the study presented here, the judges seem to be more consistent in their identification of borderline-examinees than in the study by Livingstone and Zieky. The standard deviations for the Angoff procedure were small, compared to the standard deviations found by Giraud et al. (1999/2000). In our study, the standard deviations were 3% of the total score for the performance level P, and 6% for PD. Giraud et al. found standard deviations ranging from 8% to 15% of the total score. For the borderline-group procedure, Livingstone and Zieky (1989) claim that another indicator of trustworthiness in cut-scores is when the means for the different groups of examinees are ordered. In this study the borderline-groups have means in between the two adjacent groups indicating credibility of the formation of borderline-groups.

These results, adding to the trustworthiness of the borderline-group procedure, are supplemented by results that indicate problems with the procedure. The median test results of borderline examinees from high performing classes are higher than the median test results of borderline examinees from low performing classes. In other words, a positive relationship is found between the median test results for the borderline-examinees and the median test results for the whole teaching group that those borderline-examinees belong to. These results are in accordance with the results presented by Livingstone and Zieky (1989). Teachers seem to be influenced by the performance level in their student group when they identify borderline-examinees making cut-scores dependent on the sampled groups of examinees participating in the borderline-group procedure. This result supports the claim of Hambleton et al. (2000) that cut-scores derived by examinee-centred methods are dependent on the representativeness of the sampled student groups. If the judges only teach high performance groups of examinees, there is a potential risk that the cut-score would be too high. Similarly, if the judges have only low performance groups, the cut-scores would be too low. A representative sample of groups at different

performance levels would give more trustworthy cut-scores. The influence of the overall level of each examinee student group in the borderline-group method is likely to be present among panelists in the Angoff method as well. It is plausible that judges teaching high-performing student groups will give higher Angoff estimates than judges teaching low-performance student groups. Further research is needed to substantiate this.

External evidence

External evidence is based on comparisons with external sources, e.g. results from other standard-setting methods, other measurements of the same knowledge and/or skills, and group distribution when the test is given. One external source of evidence comes from the comparison of results from the two standard-setting methods. In this study the two standard setting procedures gave similar cut-scores for one performance level (P), but different cut-scores for the other (PD). Because of the similarity in cut-scores for P these cut-scores are more trustworthy than the cut-scores for PD, following the claim made by Hambleton and Pitoniak (2006) that a cut-score is viewed as more trustworthy if different standard-setting procedures result in similar performance standards.

Another external source of evidence is the course grades that teachers give their examinees. The correlations between test-results and course grades were similar and fairly high for both procedures (0.81 – 0.83). For P, the percentage agreement between course- and performance level on the test was higher for the borderline-group method. For PD the Angoff method gave higher agreement. The differences in correlation between the two standard-setting methods were small, which makes the evidence inconclusive as to which standard-setting method is more trustworthy in the light of correlation to course-grades. Course grades can be viewed as a valid source of external evidence in the evaluation of performance standards used in national tests because they are based on the same learning objectives and are intended to be measures of the same domain. However, a single test always represents a narrower domain of learning objectives because of the restricted time for testing and because of the difficulties (and costs) of using test-formats other than pencil-and-paper-tests. Furthermore, course grades are not independent of test results since the test-result is one piece of information that teachers use for grading their examinees.

In addition, external evidence is retrieved from the comparison of Angoff estimates with the actual performances of examinees at the cut-scores arrived by using the Angoff procedure. Ideally these would coincide. However, for a number of items, the Angoff estimates are either significant over- or underestimations. These deviations seem acceptable for the cut-score for the lower performance level. However, based on our study we can conclude that the number of items with deviations was higher for the cut-score for the higher performance level. This makes the standard setting at the higher performance level more questionable using the Angoff procedure.

CONCLUSION

We conclude that both the Angoff method and the borderline-group method provide reasonable and trustworthy approaches to standard setting. Our study has exposed some of the validity issues concerning standard setting and confirmed some of the potential problems with the methods, e.g. the differences between borderline-groups identified by teachers of high- and low-performing groups. Messick (1989, p. 13) defines validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment". Standard-setting methods add to the validity arguments for a test by being based on theoretical rationales and performed according to the recommendations in the literature. However, it is also important that these rationales are supported by empirical evidence in follow-up studies. Furthermore, standard-setting methods should be evaluated from different perspectives, including aspects as cost-efficiency, comparability and long-term consequences. Further studies are needed to better understand the implications for an evidence-based practice based on sound methods for standard setting.

REFERENCES

Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development.* Mahwah: Lawrence Erlbaum Associations.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks: SAGE Publications.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centred method for setting

standards on achievement test. *Applied Measurement in Education, 12*(4), 343-366.

Ferdous, A. A., & Plake, B. S. (2007). Item selection strategy for reducing the number of items rated in an Angoff standard setting study. *Educational and Psychological Measurement, 67*(2), 193-206.

Giraud, G., Impara, J. C., & Buckendahl, C. (1999/2000). Making the cut in school districts: alternative methods for setting cutscores. *Educational Assessment, 6*(4), 291-304.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355-366.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.) *Educational Measurement* (4th ed.). Westport: American Council on Education & Praeger Publishers.

Hansche, L. N. (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I.* Washington: U.S. Department of Education & The Council of Chief State School Officers.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63*(4), 584-601.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed.). New York: American Council of Education & McMillan.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461.

Kane, M. (1998a). Choosing between examinee-centred and test-centred standard-setting methods. *Educational Assessment, 5*(3), 129-145.

Kane, M. (1998b). Criterion bias in examinee-centred standard setting: Some thought experiments. *Educational Measurement: Issues and Practice, 17*(1), 23-30.

Kane, M. T. (2001). So much remain the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.) *Setting performance standards. Concepts, methods, and perspectives* (pp. 53-88). Mahwah, N. J.: Lawrence Erlbaum Associates.

Lindström, J. O. (2003). *The Swedish national course tests in mathematics.* (EM-report no. 43). Umeå: Department of Educational Measurement, Umeå University.

Livingstone, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton: Educational Testing Services.

Livingstone, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2*(2), 121-141.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: American Council on Education.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.) *Setting performance standards. Concepts, methods, and perspectives.* Mahwah, N. J.: Lawrence Erlbaum Associates.

Skolverket. (2005). *National assessment and grading in the Swedish school system.* Stockholm: Skolverket.

Citation

Nässtrom, Gunilla and Peter Nyström (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment Research & Evaluation*, 13(9). Available online: http://pareonline.net/getvn.asp?v=13&n=9

Authors

Gunilla Näsström
Department of Educational Measurement
Umeå university
SE-90187 Umeå
Sweden
e-mail: gunilla.nasstrom [at] edmeas.umu.se

Peter Nyström
Department of Educational Measurement
Umeå university
SE-90187 Umeå
Sweden
e-mail: peter.nystrom [at] edmeas.umu.se